Week 3: Churn Analysis & Predictive Modeling Report

Executive Summary

This report presents a comprehensive churn analysis and predictive modeling exercise using the Week 1 cleaned dataset. The objectives were to:

- Develop predictive models to forecast student drop-offs (churn).

- Evaluate model performance using standard metrics.

- Identify key factors contributing to churn and propose actionable recommendations.

Table of Contents

1. Introduction

2. Data Preparation

3. Exploratory Data Analysis

4. Predictive Modeling

5. Churn Analysis

6. Recommendations

7. Conclusion

Appendix: Methods & Code References

1. Introduction

Objective and Importance of Churn Analysis

Student churn (drop-out) is a critical metric that impacts program effectiveness, resource allocation, and overall learning outcomes. Early identification of at-risk students allows institutions to intervene proactively, improving retention and completion rates.

This report defines churn as instances where a student is recorded with Status Description either 'Withdraw' or 'Dropped Out'. The analysis explores patterns, builds prediction models, and provides recommendations tied to operational actions.

Scope and Learning Outcomes

The analysis covers predictive modeling, churn factor identification, and recommended intervention strategies. Learning outcomes include developing predictive modeling skills, gaining expertise in churn analysis, and practicing effective report writing.

2. Data Preparation

Data Source:
The input dataset 'Week 1 Deliverable - Data Cleanup (1).xlsx' contained 8,560 records and columns representing demographic, application, and program details.

Cleaning Steps:
- Filtered records to keep meaningful statuses.
- Created binary churn label: churn=1 for 'Withdraw' and 'Dropped Out'.
- Parsed date fields and engineered features: Age_at_Apply, Apply_to_Start_Days, Start_to_End_Days, Signup_to_Apply_Days.
- Handled missing values by imputation (median for numeric, most frequent for categorical) during preprocessing.

Split:
Train/test split used a 75/25 stratified allocation to preserve churn distribution.

Data Summary

Records (rows)	8558
Features (columns)	23
Churn positive rate	8.21%
Train/Test split	75% train / 25% test (stratified)

3. Exploratory Data Analysis

This section highlights distributions and relationships observed in the cleaned dataset. Figures below show churn distribution, age distribution, and churn by country (top 10).

Figure 1: Churn distribution (counts).

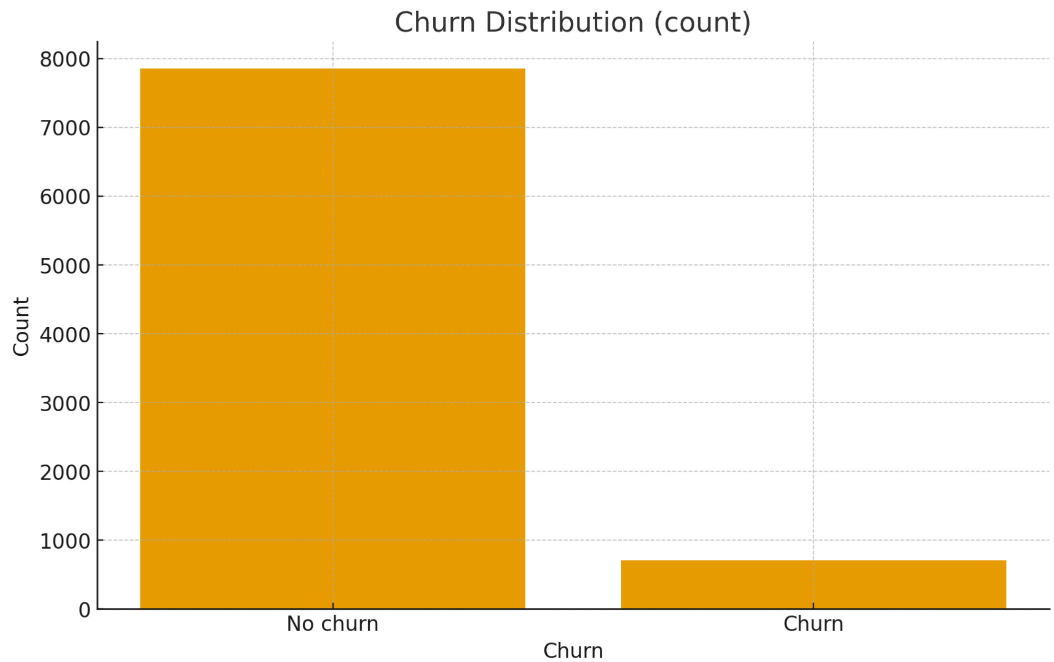


Figure 2: Age at application (histogram).

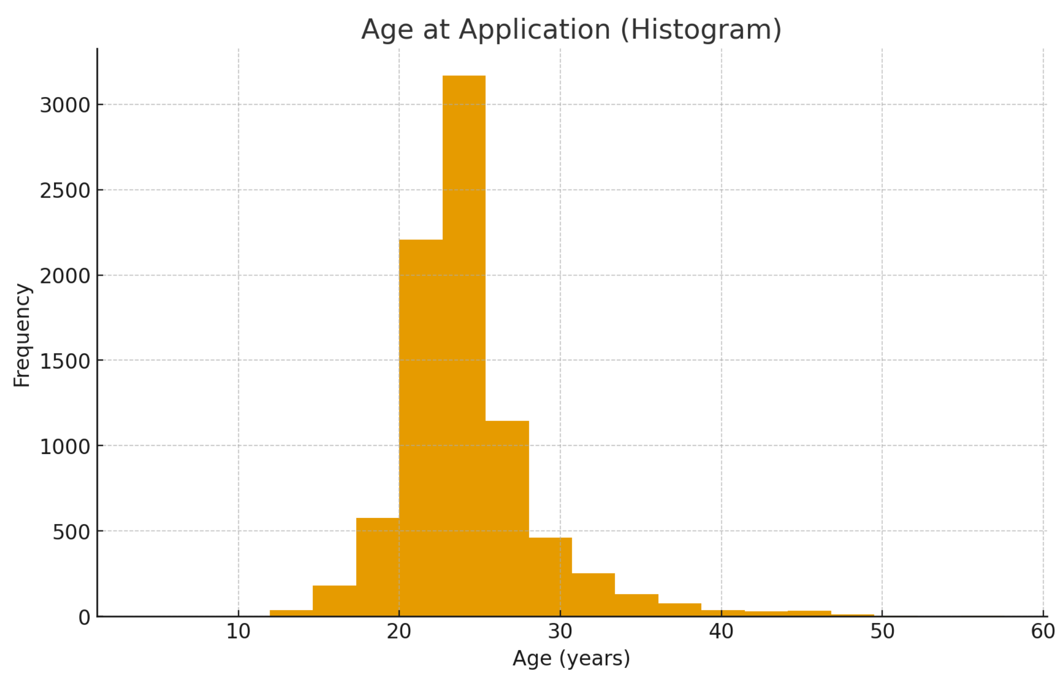
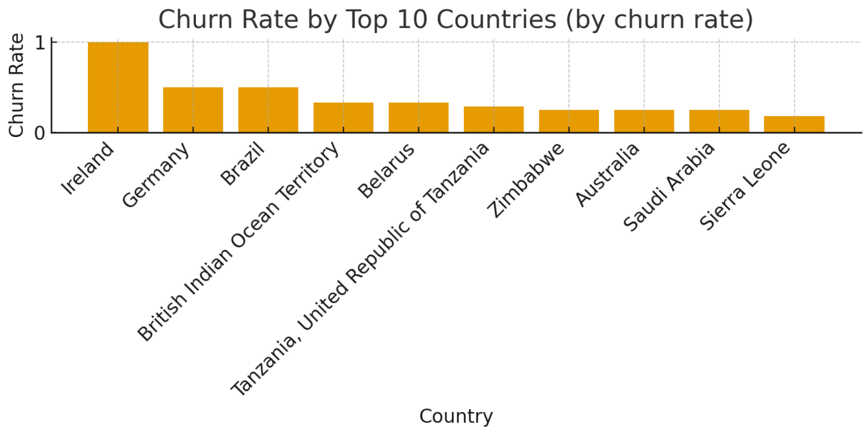


Figure 3: Churn rate by top countries (by churn rate).



4. Predictive Modeling

Models trained: Logistic Regression, Random Forest, Gradient Boosting. Models use one-hot encoding for categoricals and scaling for numeric features. Evaluation metrics are shown in Table 1.

Model	Accuracy	Precision	Recall	F1	ROC_AUC
Gradient Boosting	0.975	0.913	0.773	0.837	0.973
Logistic Regression	0.962	0.701	0.932	0.800	0.972
Random Forest	0.974	0.870	0.801	0.834	0.966

Figure 4: ROC Curves for the trained models.

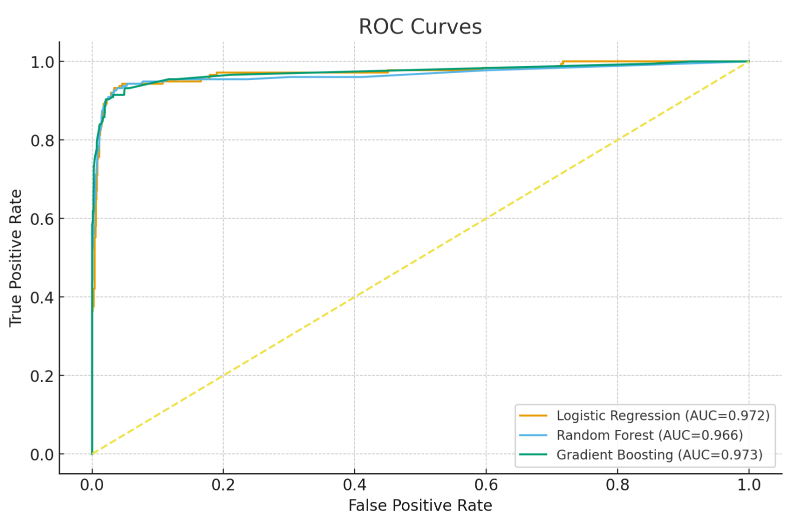
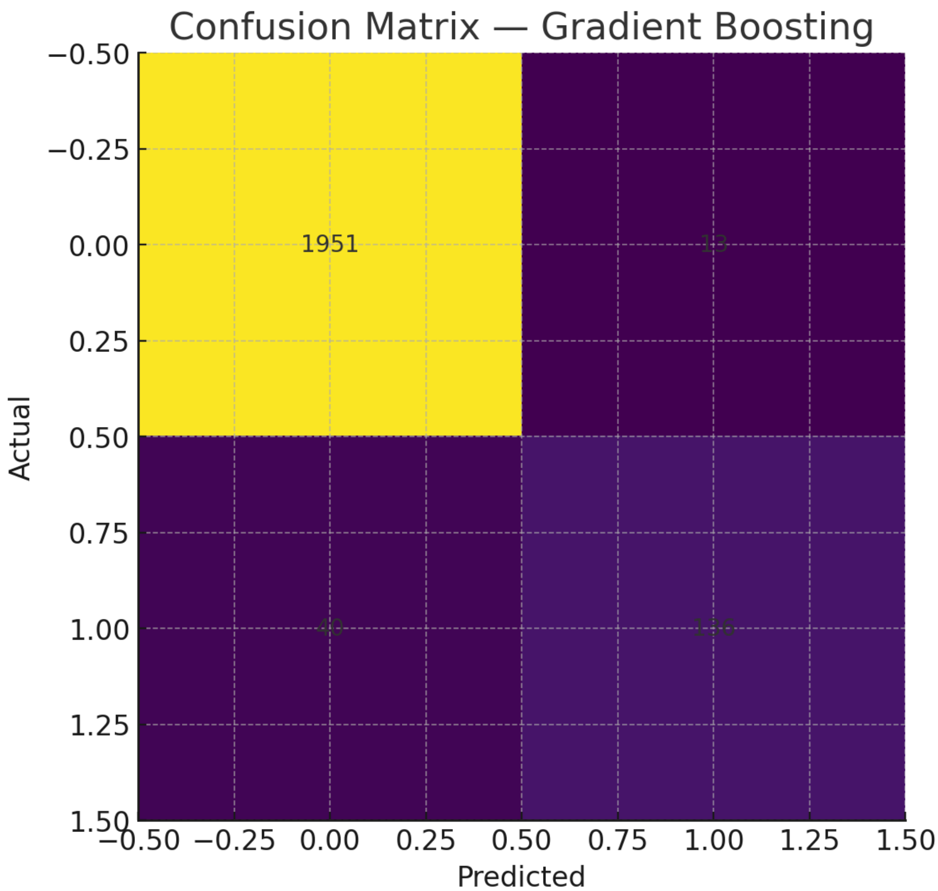


Figure 5: Confusion

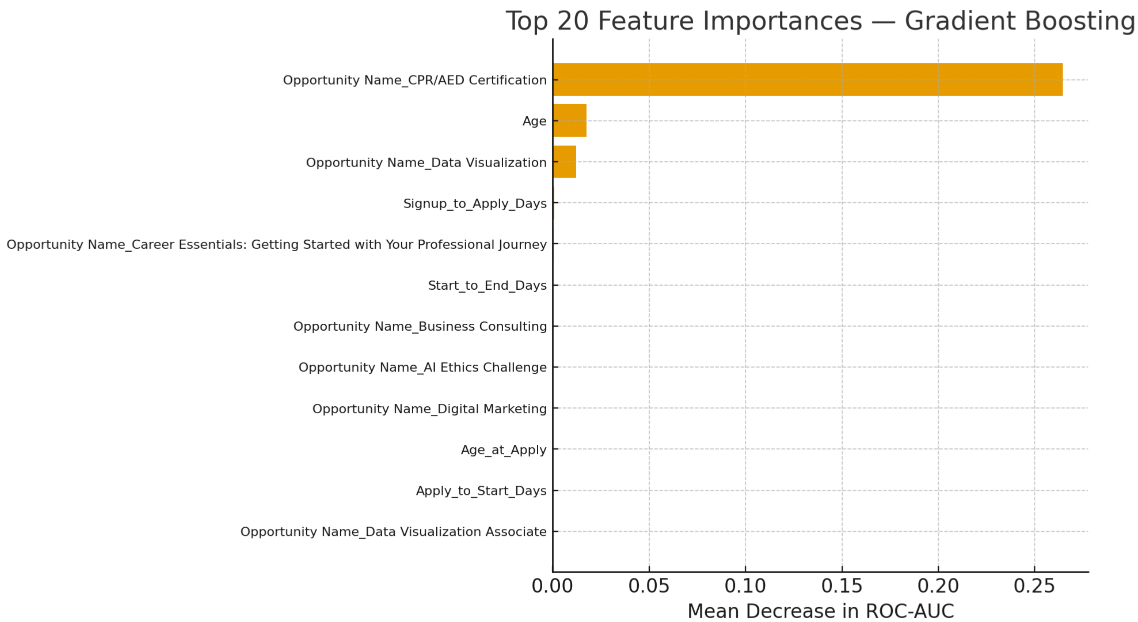


Matrix for the best model (Gradient Boosting).

5. Churn Analysis

Feature importance (permutation importance on test set) identifies the factors that most influence churn predictions. Figure 6 shows the top features.

Figure 6: Top 20 Feature Importances



Interpretation of Top Drivers

The following features were among the most influential in predicting churn:

- Opportunity Name_CPR/AED Certification

- Age

- Opportunity Name_Data Visualization

- Signup_to_Apply_Days

- Opportunity Name_Career Essentials: Getting Started with Your Professional Journey

- Start_to_End_Days

- Opportunity Name_Business Consulting

- Opportunity Name_AI Ethics Challenge

- Opportunity Name_Digital Marketing

- Age_at_Apply

6. Recommendations

- Prioritize early outreach for applicants with longer Apply_to_Start_Days to reduce waiting-time churn.

- Implement automated reminders and nudges for applicants who show long Signup_to_Apply_Days.

- Target high-risk Opportunity Categories and geographic segments with tailored support.

- Create a weekly 'risk list' from model probabilities for advisors to follow up.

- Offer onboarding micro-sessions and mentorship in the first two weeks to stabilize new cohorts.

- A/B test different communication cadences for flagged high-risk students to find the most effective interventions.

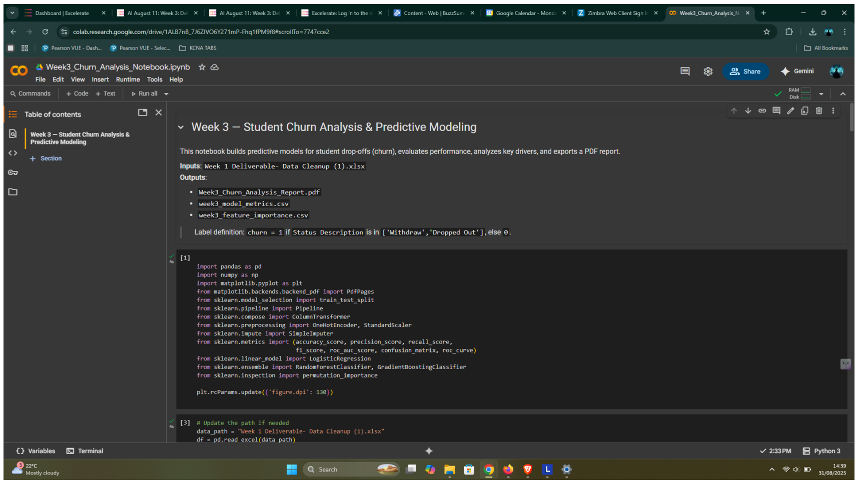
7. Conclusion

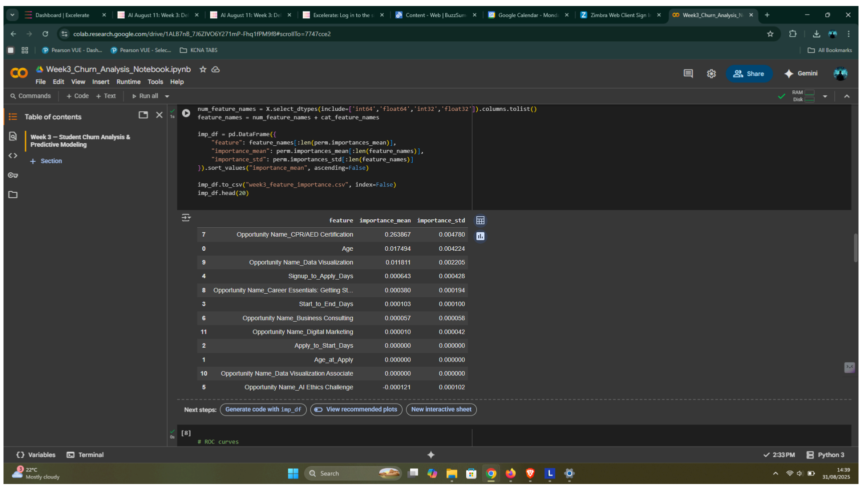
This analysis demonstrates that predictive modeling can reliably identify students at risk of dropping out. By operationalizing the model outputs and focusing on the features most strongly associated with churn, the program can deploy targeted interventions and improve retention outcomes.

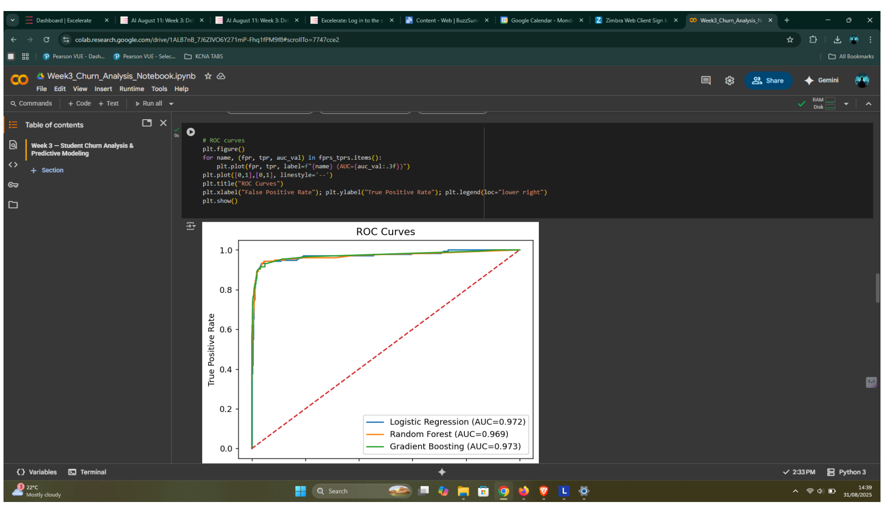
Appendix: Methods & Code References

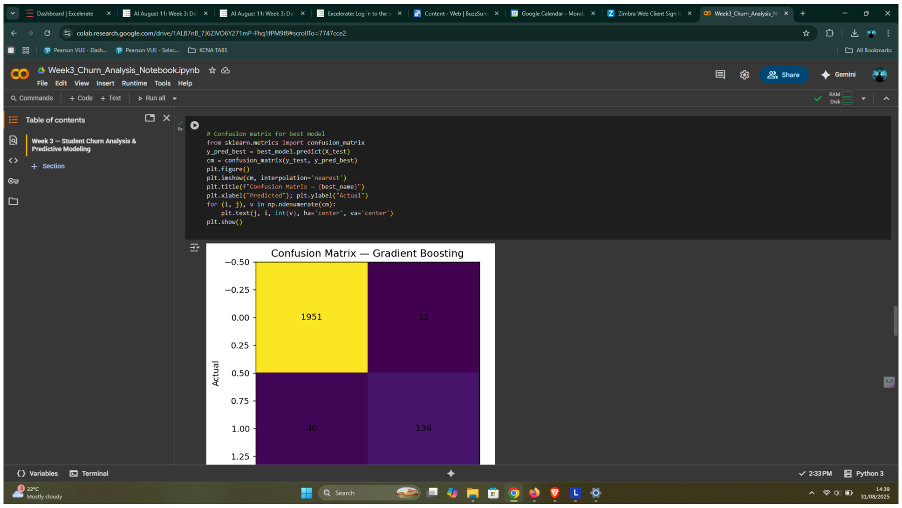
Key implementation notes:
- Categorical encoding: One-Hot
- Numeric imputation: median
- Models: Logistic Regression (balanced), Random Forest (balanced), Gradient Boosting
- Evaluation: Accuracy, Precision, Recall, F1, ROC-AUC

Code and Jupyter notebook are included as separate deliverables and reproduce all steps used to build this report.

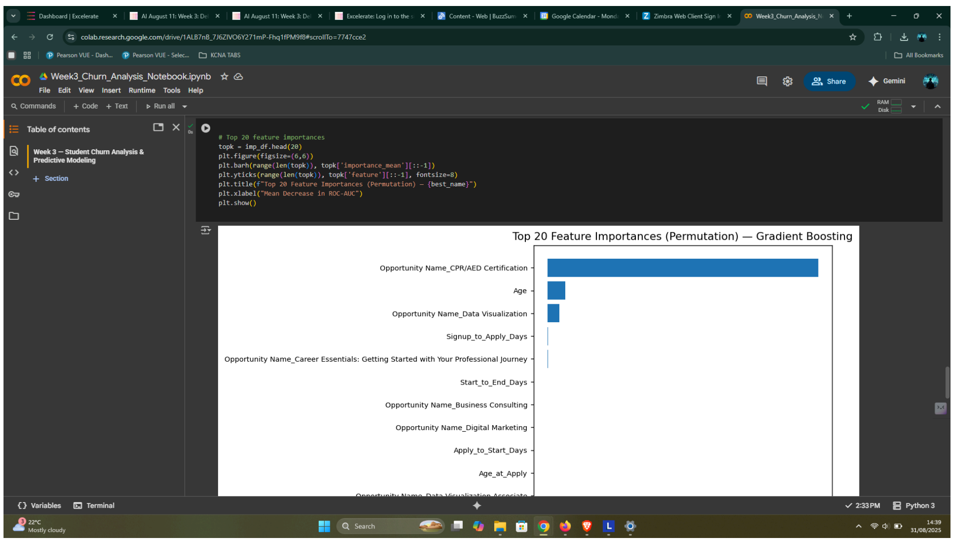

Week3_Churn_Analysis_Notebook.ipynb
Table of contents
Week 3 — Student Churn Analysis & Predictive Modeling
Week 3 — Student Churn Analysis & Predictive Modeling
This notebook builds predictive models for student drop-offs (churn), evaluates performance, analyzes key drivers, and exports a PDF report.
Inputs: Week 1 Deliverable- Data Cleanup (1).xlsx
Outputs:
Week3_Churn_Analysis_Report.pdf
week3_model_metrics.csv
week3_feature_importance.csv
Label definition: churn = 1 if Status Description is in ['Withdraw','Dropped Out'], else 0.
[1]
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.backends.backend_pdf import PdfPages
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.metrics import (accuracy_score, precision_score, recall_score,
f1_score, roc_auc_score, confusion_matrix, roc_curve)
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn.inspection import permutation_importance
plt.rcParams.update({'figure.dpi': 130})
[3]
# Update the path if needed
data_path = "Week 1 Deliverable- Data Cleanup (1).xlsx"
df = pd.read_excel(data_path)
2:33 PM
Python 3
14:39
31/08/2025


Week3_Churn_Analysis_Notebook.ipynb
num_feature_names = X.select_dtypes(include=['int64','float64','int32','float32']).columns.tolist()
feature_names = num_feature_names + cat_feature_names
imp_df = pd.DataFrame({
"feature": feature_names[:len(perm.importances_mean)],
"importance_mean": perm.importances_mean[:len(feature_names)],
"importance_std": perm.importances_std[:len(feature_names)]
}).sort_values("importance_mean", ascending=False)
imp_df.to_csv("week3_feature_importance.csv", index=False)
imp_df.head(20)
feature importance_mean importance_std
7 Opportunity Name_CPR/AED Certification 0.263867 0.004780
0 Age 0.017494 0.004224
9 Opportunity Name_Data Visualization 0.011811 0.002205
4 Signup_to_Apply_Days 0.000643 0.000428
8 Opportunity Name_Career Essentials: Getting St... 0.000380 0.000194
3 Start_to_End_Days 0.000103 0.000100
6 Opportunity Name_Business Consulting 0.000057 0.000058
11 Opportunity Name_Digital Marketing 0.000010 0.000042
2 Apply_to_Start_Days 0.000000 0.000000
1 Age_at_Apply 0.000000 0.000000
10 Opportunity Name_Data Visualization Associate 0.000000 0.000000
5 Opportunity Name_AI Ethics Challenge -0.000121 0.000102
Next steps: Generate code with imp_df View recommended plots New interactive sheet
[8]
ROC curves
2:33 PM
Python 3
14:39
31/08/2025


Week3_Churn_Analysis_Notebook.ipynb
# ROC curves
plt.figure()
for name, (fpr, tpr, auc_val) in fprs_tprs.items():
plt.plot(fpr, tpr, label=f"{name} (AUC={auc_val:.3f})")
plt.plot([0,1],[0,1], linestyle='--')
plt.title("ROC Curves")
plt.xlabel("False Positive Rate"); plt.ylabel("True Positive Rate"); plt.legend(loc="lower right")
plt.show()
ROC Curves
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
Logistic Regression (AUC=0.972)
Random Forest (AUC=0.969)
Gradient Boosting (AUC=0.973)
2:33 PM
Python 3
14:39
31/08/2025



Week3_Churn_Analysis_Notebook.ipynb
Table of contents
Week 3 — Student Churn Analysis & Predictive Modeling
# Confusion matrix for best model
from sklearn.metrics import confusion_matrix
y_pred_best = best_model.predict(X_test)
cm = confusion_matrix(y_test, y_pred_best)
plt.figure()
plt.imshow(cm, interpolation='nearest')
plt.title(f"Confusion Matrix — {best_name}")
plt.xlabel("Predicted"); plt.ylabel("Actual")
for (i, j), v in np.ndenumerate(cm):
    plt.text(j, i, int(v), ha='center', va='center')
plt.show()
Confusion Matrix — Gradient Boosting
Actual
1951
13
40
136



Week3_Churn_Analysis_Notebook.ipynb
Table of contents
Week 3 — Student Churn Analysis & Predictive Modeling
# Top 20 feature importances
topk = imp_df.head(20)
plt.figure(figsize=(6,6))
plt.barh(range(len(topk)), topk['importance_mean'][::-1])
plt.yticks(range(len(topk)), topk['feature'][::-1], fontsize=8)
plt.title(f"Top 20 Feature Importances (Permutation) — {best_name}")
plt.xlabel("Mean Decrease in ROC-AUC")
plt.show()
Top 20 Feature Importances (Permutation) — Gradient Boosting
Opportunity Name_CPR/AED Certification
Age
Opportunity Name_Data Visualization
Signup_to_Apply_Days
Opportunity Name_Career Essentials: Getting Started with Your Professional Journey
Start_to_End_Days
Opportunity Name_Business Consulting
Opportunity Name_Digital Marketing
Apply_to_Start_Days
Age_at_Apply